Links between Manual Punctuation Marks and Automatically Detected Prosodic Structures

Katarina Bartkova[1], *Denis Jouvet*[2]

[1] ATILF - Analyse et Traitement Informatique de la Langue Française
Université de Lorraine, ATILF, UMR 7118, Nancy, F-54063, France
[2] Speech Group, LORIA
Inria, Villers-lès-Nancy, F-54600, France
Université de Lorraine, LORIA, UMR 7503, Villers-lès-Nancy, F-54600, France
CNRS, LORIA, UMR 7503, Villers-lès-Nancy, F-54600, France
katarina.bartkova@atilf.fr, denis.jouvet@loria.fr

Abstract

This paper presents a study of the links between punctuation and automatically detected prosodic structures, as observed on large speech corpora that were manually annotated during French speech transcription evaluation campaigns. These corpora contain more than 3 million words and almost 350 thousands punctuation marks. The detection of the prosodic boundaries and of the prosodic structures is based on an automatic approach that integrates little linguistic knowledge and mainly uses the amplitude and the direction of the F0 slopes as described in [1], as well as phone durations. The paper first analyzes the occurrences of the punctuation marks with respect to various sub-corpora, which also highlights the variability among annotators. Then, we focus on analyzing prosodic parameters with respect to the punctuation marks, followed or not by a pause, and on analyzing the links between the automatically detected prosodic structures and the manually annotated punctuation marks.

Index Terms: prosodic structure, speech, punctuation

1. Introduction

Speech is structured by prosodic means to allow the listener to access to lexical units and therefore to the meaning conveyed by the speech signal. For optimal results, most of the automatic speech processing techniques (automatic translation, information retrieval…) need to recover the speech prosodic structuring. This corresponds to adding punctuation marks to the raw streams of words supplied by automatic speech transcription systems. Though orthographic conventions used to capture the speech prosody cannot reflect all the various linguistic and extra-linguistic meanings of the speech prosody, however, they allow to mark its most elementary linguistic functions such as the modality and the finality of a sentence through full stops, exclamation marks, question marks… and through commas the intention of the speaker after a deep prosodic boundary (prosodic group inserted closer to the root than to the leaves in the prosodic tree, cf. Figure 1) to continue to develop the same sentence.

The pattern of the prosodic parameters (mainly F0 movement and syllable duration lengthening) depends on the type of the prosodic boundary. A continuation (major – deeper boundary; or minor – shallower boundary) is indicated by the slope of the F0 movement whose direction can be rising or falling. The end of a declarative sentence is indicated by a falling F0 movement [2],[1] although other (flat or rising) movements are also observed in French spontaneous speech [3]. The modality of the sentence is expressed by the direction and the steepness of the F0 movement: polar (yes-no) questions in French are marked only by a sharply rising F0 slope while an order is marked by a sharply falling F0 slope.

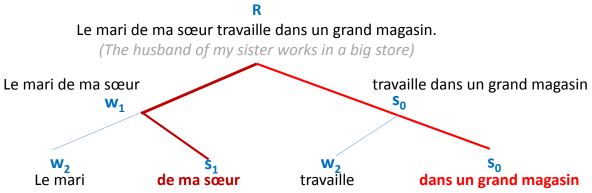
R
Le mari de ma sœur travaille dans un grand magasin.
(The husband of my sister works in a big store)
Le mari de ma sœur
w1
travaille dans un grand magasin
s0
w2
Le mari
s1
de ma sœur
w2
travaille
s0
dans un grand magasin

Figure 1: *Metric tree with the deepest prosodic boundaries at nodes s_1 and s_0 (little cohesion to the right)*

In French the syllable duration is closely related to the prosodic boundaries: the duration of a syllable on a continuation prosodic boundary is significantly lengthened compared to the duration of syllables on non-prosodic boundaries. Moreover, the syllable duration, although lengthened on sentence final position, is however shorter than on clause final (major or minor continuation) position [4].

The automatic detection of the prosodic structure used in this study, is based on a theoretical description of prosodic trees; the framework was first developed for prepared speech [1] and later adapted for semi-spontaneous speech [5]. The approach was recently revisited and applied on various types of speech material, including spontaneous speech [6].

In this study, the location of the manually set punctuation marks is compared with prosodic units detected by the automatic approach described in [6]. This comparison allows us observing whether human annotators were or were not influenced by the prosodic parameter values considered as pertinent for the automatic segmentation and also if there is coherence between the location of the punctuation marks and the depth of the prosodic boundaries estimated from their prosodic parameters.

The paper is organized as follows. Section 2 presents the speech corpora used as well as some global statistics about the punctuation marks observed in the manually transcribed data. Section 3 summarizes the process used to obtain automatically the prosodic structures of the speech data. Then, Section 4 analyzes the relations between the punctuation marks and the prosodic parameters, the prosodic groups and the prosodic structures.

10.21437/SpeechProsody.2014-67

2. Speech corpora and punctuation

Several French speech corpora have been used in this study. They come from the recent ESTER2 [7] and ETAPE [8] speech transcription evaluation campaigns, and from the EPAC project [9],[10]. The ESTER2 data are French broadcast news collected from various radio channels; thus mainly prepared speech, plus some interviews. The EPAC data are mainly spontaneous speech, recorded during the ESTER1 campaign, and correspond to shows from three French radios [9]. The ETAPE data corresponds to debates collected from radio and TV channels, and is mainly spontaneous speech. The training part of the ESTER2 and ETAPE data, plus the transcribed part of the EPAC corpus correspond to a total of almost 300 hours of signal and 4 million running words. The development and test parts of the corpora have been left aside for further experiments.

The values of F0 in semitones and of the energy are computed every 10 ms from the speech signal using the ETSI/AURORA [11] acoustic analysis. The phonetic transcription of the text, with pronunciation variants, is obtained using the BDLEX [12] lexicon and an automatic grapheme-to-phoneme transcription system [13] which is applied for words absent from the lexicon. The forced text-speech alignment is carried out with the Sphinx tools [14]. This provides the speech segmentation into phonemes and words, which is then used to compute the sound durations, as well as to obtain the location and the duration of the pauses. As the speech signal quality is rather good, it can be assumed that the segmentation is obtained out without major problems. After the forced alignment step, short pauses of less than 100 ms occurring before a plosive or a fricative are removed. Finally, end of speech events, as well as last word before a speaker change, falls in the columns “plus pause” in the following tables.

Punctuation was present in most of the manual transcription files, which correspond to more than 3 million words (lexical and grammatical words). Table 1 exhibits some global statistics for each of the three corpora. There is on average one punctuation symbol every 8 to 10 words (every 8.3 words for EPAC, every 9.8 words for ETAPE).

Table 1. *Size of the speech corpora used, and number of words and punctuation symbols with respect to position (followed or not by a pause).*

	Files	Words			Punctuation symbols		
		count	plus pause	no pause	count	plus pause	no pause
ESTER	292	2.01 M	18.2%	81.8%	225 k	66.7%	33.3%
EPAC	106	0.92 M	18.5%	81.5%	94 k	50.9%	49.1%
ETAPE	44	0.24 M	20.8%	79.2%	29 k	61.9%	38.1%

With respect to pauses, the three corpora have a similar behavior, on average a pause is observed after every 5 words (about 20% of the words are followed by a pause). However there are more differences for what concern manually set punctuation symbols. About one third of the punctuation symbols are not followed by a pause for the ESTER and ETAPE corpora, whereas this is almost one punctuation symbol out of two which is not followed by a pause in the EPAC corpus. This exhibits a rather large variation in punctuation annotation which is due to annotators as the high frequency of punctuation symbols not followed by a pause in the EPAC corpus does not match neither with the annotations on ESTER (similar radios) nor with the annotations on ETAPE (also mainly spontaneous speech).

Table 2. *Analysis of the various punctuation marks in the three speech corpora.*

	dot	excl.	inter.	3 dots	semi col.	two dots	comma
ESTER	32.9%	1.8%	2.3%	0.8%	1.3%	2.6%	58.3%
EPAC	23.0%	1.6%	4.2%	0.0%	5.9%	5.5%	59.8%
ETAPE	32.0%	2.0%	5.1%	0.5%	0.2%	2.3%	57.9%

Table 2 presents the distribution of the various punctuation marks in the 3 corpora. The frequency of the comma is rather similar between the 3 corpora. However, other punctuation symbols on EPAC exhibit a different behavior, especially with respect to semi-column and two dots marks which are much more used by the annotators of the EPAC corpus than by the annotators of the ESTER and ETAPE corpus.

3. Automatic detection of prosodic structures

As mentioned before, the approach used for the automatic detection of prosodic structures is based mainly on prosodic parameter values and little linguistic knowledge. The process starts by an initial segmentation of the text (i.e., the sequence of words corresponding to the speech signal) into potentially stressed prosodic units. This is carried out by grouping grammatical words with lexical words. Prosodic parameters are then considered only on the vowels of the last syllables of the potentially stressed groups. Two main parameters, the F0 slope and the normalized duration of the vowels in final positions (other than the schwa vowel in final position when the word is plurisyllabic) are used to detect prosodic boundaries. The duration threshold that separates stressed vowels from unstressed vowels was determined from the analysis of the distribution of the duration of vowels in unstressed positions (only syllables other than last syllables of the lexical units were considered for this estimation) and in stressed positions (only syllables followed by a pause were considered for this estimation). A similar approach was applied for determining the threshold that separates the values of the slopes of F0 on prosodic and non-prosodic boundaries. A third parameter, the variation of the F0 value (obtained as the difference in F0 between the current vowel and the previous one not separated from the current vowel by a pause) is also calculated for the last vowel in the prosodic groups. If this variation of the F0 value is higher than 5 semi-tones then a prosodic boundary is set on this syllable.

The automatic approach also evaluates the depth of the prosodic boundaries. A prosodic boundary which is marked by a steep F0 slope (higher than the glissando threshold for speech) and a long vowel duration, or a very long vowel duration and a more moderate F0 slope (higher than the glissando threshold for vowels), receives the symbolic annotation C0. To capture smaller variations of the prosodic parameters, symbolic annotation ranging from C1 (deeper prosodic boundary) to C5 (shallower prosodic boundary) are used. To avoid a too fine-grained prosodic segmentation, prosodic boundaries whose symbolic annotation is C3 and whose length is less than 2 syllables, are neutralized and attached to the following prosodic group. Also, when the prosodic group exceeds 10 syllables, an intermediate prosodic boundary is searched around the middle part of the group using this time lower threshold values for vowel duration and F0 slope detection. When an appropriate split is found, the prosodic group is cut into 2 groups; otherwise the group regardless of its length is maintained as one single prosodic group. The symbolic annotations (C0, C1, ...) are used to construct prosodic trees for each breath group (speech signal preceeded and followed by a long pause). In the prosodic tree construction, a prosodic group is attached to the next prosodic group if the next group has a lower symbolic depth (i.e. if the next group is closer to the root of the prosodic tree).

4. Prosodic structure and punctuation

This section is dedicated to analyzing in details, on the ESTER training data, the punctuation marks with respect to the prosodic parameters, the prosodic groups and the prosodic structures that were automatically obtained. Besides Table 3, most of the analyzed items are presented as normalized frequency histograms; that means that the figures show the distribution of some parameters (e.g., pause duration, F0 slope …) with respect to the presence or absence of punctuation symbols at the end of prosodic groups. To keep the figures simple, only the dot and the comma punctuation symbols are considered, plus the "*no punctuation*" case, corresponding to end of prosodic groups that are not associated to any punctuation symbol. Moreover, two histograms are drawn in each case, whether a pause follows or not.

4.1. Punctuation and prosodic groups

The first analysis concerns the position of the punctuation marks with respect to the automatically obtained prosodic groups. Table 3 shows that most of the punctuation marks that are followed by a pause (column *plus pause*) match with the end of automatically detected prosodic groups. Moreover, almost two thirds of the punctuation marks that are not followed by a pause also match with the end of prosodic groups. Table 3 shows that, overall, less than 14% of the punctuation marks fall inside the automatically detected prosodic groups. For these cases, if we left aside the few cases were the detection of the prosodic groups is not correct, the annotator's decision was probably influenced more by semantic or syntactic information not marked by prosodic parameters.

Table 3. *Analysis of the various punctuation marks with respect to position (any place or end of automatically detected prosodic groups).*

ESTER		Any place		End prosodic group	
		plus pause	no pause	plus pause	no pause
dot	74154	89.0%	11.0%	89.0%	7.3%
excl.	4026	89.1%	10.9%	88.1%	5.9%
inter.	5163	86.6%	13.4%	86.6%	8.6%
3 dots	1742	84.3%	15.7%	83.5%	9.3%
semi-col.	3018	81.3%	18.7%	80.1%	12.4%
two dots	5969	65.8%	34.2%	63.3%	22.6%
comma	131280	52.1%	47.9%	50.3%	29.7%
Total	*225352*	*66.7%*	*33.3%*	*65.5%*	*20.8%*

For what concerns the length of the prosodic groups, Figure 2 shows that prosodic groups reduced to a single word are almost never followed by a punctuation mark. Prosodic groups followed by a punctuation mark are longer than those not associated to any punctuation mark. Also, the distributions of the length of the prosodic groups followed by a dot or by a comma are very similar. Moreover, the distribution of the length of the prosodic groups is not significantly different whether the prosodic groups are followed by a pause or not.

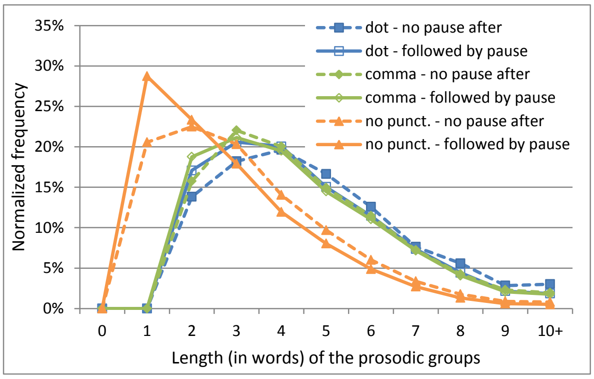

Figure 2: *Normalized frequency histograms of the length of the prosodic groups*

Figure 3 presents the distributions of the pause durations when observed after a prosodic group. When there is no punctuation associated to the prosodic group, the duration of the following pause is most of the time smaller than 200 ms. When a punctuation mark is associated to the prosodic group, the duration of the following pause if typically between 300 ms and 500 ms. Overall, the duration of the pause following a prosodic group tends to be longer for prosodic groups associated to dots than for prosodic groups associated to commas, which are themselves longer than for prosodic groups that are not associated to any punctuation mark.

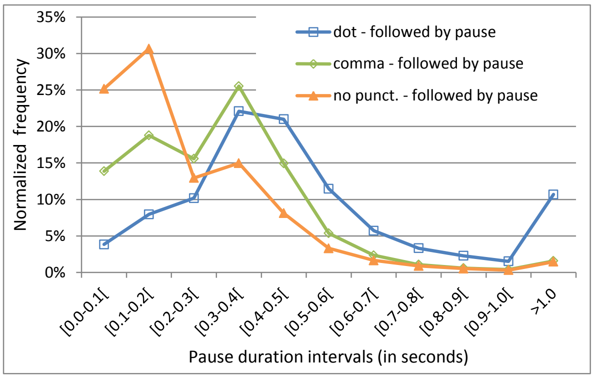

Figure 3: *Normalized frequency histograms of the duration of the pauses after prosodic groups.*

4.2. Punctuation and prosodic parameters

The first prosodic parameter considered here is associated to the global F0 slope, that is the longest F0 slope that ends in the last syllable of the prosodic group. Figure 4 presents the normalized histograms of the absolute variation of F0 (delta F0) between the beginning and the end of the global slope. The delta F0 (Figure 4) and the F0 slope (not represented here) are slightly higher for prosodic groups associated to the dot and comma punctuation marks than for prosodic groups not associated to any punctuation mark. That means that this parameter is not perceived as pertinent by human annotators for the decision of the punctuation marks.

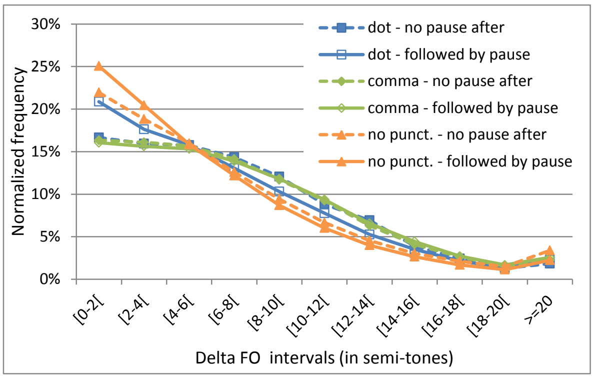

Figure 4: *Normalized frequency histograms of the delta F0 absolute values of the global slope ending in each prosodic group.*

The analysis of the last F0 value on the prosodic groups with respect to the associated punctuation mark is reported in Figure 5. The F0 value was first normalized according to the speaker F0 range, and thus expressed as a percentage of the F0 speaker range (0.0 meaning the lowest F0 value, 1.0 meaning the highest F0 values for the given speaker). It appears from the figure that the highest last F0 values are frequently associated to the comma punctuation mark, intermediate last F0 values are generally not associated to any punctuation mark, whereas dots are associated either to low last F0 values when the prosodic group is followed by a pause or to high last F0 values when there is no following pause.

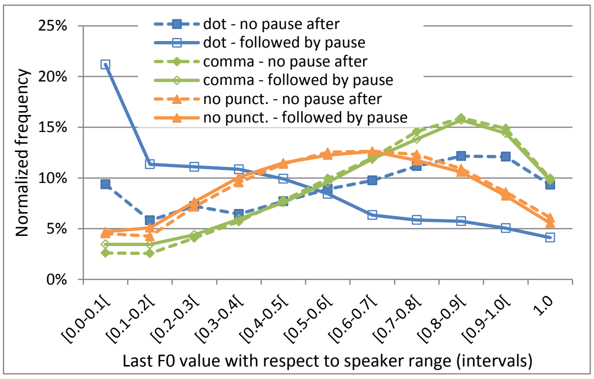

Figure 5: *Normalized frequency histograms of the F0 values at the end of each prosodic group (F0 value expressed as a ratio of the F0 speaker range).*

4.3. Punctuation and prosodic structure

Figure 6 illustrates the relation between the punctuation marks and the level of the prosodic group in the prosodic structures detected automatically. The figure shows that when there is no following pause, the dot and comma are almost always associated to prosodic groups of level 0, i.e., the prosodic group with the deepest boundary in the considered speech segment. However, many prosodic groups of level 0 are also associated to the no punctuation case. When there is no following pause, dot and comma punctuation marks are frequently associated to prosodic groups of level 1. Higher level prosodic groups (level 2 or more) are more frequently observed for the no punctuation case.

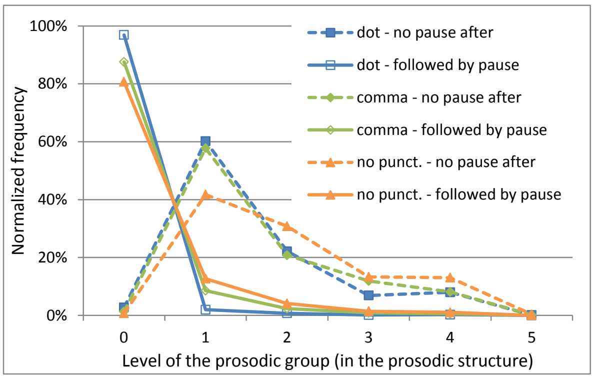

Figure 6: *Normalized frequency histograms of the level of the prosodic group (level in the associated prosodic structure).*

A more detailed analysis shows that for the no punctuation case and for the prosodic groups of level 1, there is a tendency to relate the punctuation symbol to the level of the following prosodic group: a dot if the following group is of level 2 or more, a comma if the following group is also of level 1, and no punctuation if the following group is of level 0.

5. Conclusions

This paper has analyzed the links between the punctuation marks and the prosodic parameters and prosodic structures of the speech data. The analysis was conducted on French speech corpora that were manually transcribed and annotated for speech transcription evaluation campaigns. Several hundred hours of signal were considered. A first statistical analysis of the punctuation marks on several sub corpora showed that the annotation of the punctuation bears a rather large variability due to human annotators.

The prosodic structures of the speech data were obtained automatically through an approach that integrates little linguistic knowledge; the approach mainly relies on the amplitude of the F0 slopes, as well as on phone durations. Most of the manually set punctuation marks match with the end of automatically detected prosodic groups (few punctuation marks fall inside automatically detected prosodic groups).

Prosodic groups, prosodic parameters and prosodic structures were also analyzed with respect to the presence or absence of punctuation marks, whether they are followed or not by a pause. Two punctuation marks were particularly studied: dot and comma. Parameters were analyzed through normalized frequency histograms revealing different behaviors for dot, comma and no punctuation case occurrences.

However, the distribution of the parameters still largely overlap, and each of the prosodic parameters cannot be used alone to decide on the punctuation (if any) that should be associated to the end of a prosodic group. Further studies will investigate the application of automatic classifiers that could handle simultaneously all the parameters for deciding on the presence and on the type of punctuation mark at the end of a prosodic groups.

6. References

[1] Martin, P.: “Prosodic and rhythmic structures in French”. *Linguistics* 25, pp. 925–949, 1987.

[2] Delattre, P.: “Les dix intonations de base du français”. *The French Review* 40 (1), pp. 1-14, 1966.

[3] Avanzi, M., Martin, P.: "L'intonème conclusif : une fin (de phrase) en soi ?". *Nouveaux cahiers de linguistique française*, 28. pp. 247-258, 2007.

[4] Bartkova K., Sorin, C. “A model of segmental duration for speech synthesis in French”. *Speech Communication* 6 (3), pp. 245-260, 1987.

[5] Segal, N., Bartkova, K.: “Prosodic structure representation for boundary detection in spontaneous French”. In *Proc. ICPhS 2007*, Saarbrücken, Germany, pp. 1197–1200, 2007.

[6] Bartkova, K., Jouvet, D.: “Automatic Detection of the Prosodic Structures of Speech Utterances”. In *Proc. SPECOM 2013*, Pilsen, Czech Republic, pp. 1-8, 2013.

[7] Galliano, S., Gravier, G., Chaubard, L.: “The Ester 2 evaluation campaign for rich transcription of French broadcasts”. In *Proc. INTERSPEECH 2009*, Brighton, UK, pp. 2583–2586, 2009.

[8] Gravier, G., Adda, G., Paulsson, N., Carr, M., Giraudel, A., Galibert, O.: “The ETAPE corpus for the evaluation of speech-based TV content processing in the French language”. In *Proc. LREC 2012*, Istanbul, Turkey, 2012.

[9] Estève, Y., Bazillon, T., Antoine, J.-Y., Béchet, F., Farinas, J.: “The EPAC corpus: Manual and automatic annotations of conversational speech in French broadcast news”. In *Proc. LREC 2010, European Conf. on Language Resources and Evaluation*, Valetta, Malta, 2010.

[10] Corpus EPAC: Transcriptions orthographiques. Catalogue ELRA, reference ELRA-S0305, http://catalog.elra.info.

[11] Speech Processing, Transmission and Quality Aspects (STQ); Distributed speech recognition; extended advanced front-end feature extraction algorithm; compression Algorithms, ETSI ES 202 212, 2005.

[12] de Calmès, M., Pérennou, G.: “BDLEX: a Lexicon for Spoken and Written French”. In *Proc. LREC 1998*, Grenade, pp. 1129–1136, 1998.

[13] Jouvet, D., Fohr, D., Illina, I.: “Evaluating grapheme-to-phoneme converters in automatic speech recognition context”. In *Proc. ICASSP 2012*, Kyoto, Japan, pp. 4821–4824, 2012.

[14] Sphinx (2011), http://cmusphinx.sourceforge.net/